a. ??? we can either use mapreduce or pig scripts to generate sample data sets. It is faster to test locally because pig script can be executed in grunt shell without executing mapreduce jobs in Hadoop.

b.

```
hadoop fs -cat dualcore/test2/part-m-00000 | head -100 > test_ad_data.txt
```

c.

(diskcentral.example.com,68)
(megawave.example.com,96)
(megasource.example.com,100)
(salestiger.example.com,141)

d.

(bassoonenthusiast.example.com,1246)
(grillingtips.example.com,4800)
(footwear.example.com,4898)
(coffeenews.example.com,5106)

Script

```
-- TODO (A): Replace 'FIXME' to load the test_ad_data.txt file.
--data = LOAD 'FIXME'
--data = LOAD 'test_ad_data.txt'
data = LOAD 'dualcore/ad_data[1-2]/part*'
                AS (
                        campaign_id:chararray,
                        date:chararray,
-- TODO (A): Replace 'FIXME' to load the test_ad_data.txt file.
--data = LOAD 'FIXME'
--data = LOAD 'test_ad_data.txt'
data = LOAD 'dualcore/ad_data[1-2]/part*'
                AS (
                        campaign_id:chararray,
                        date:chararray,
                        time:chararray,
                        keyword:chararray,
                        display_site:chararray,
                        placement:chararray,
                        was_clicked:int,
                        cpc:int
```

```
                );

-- TODO (B): Include only records where was_clicked has a value of 1
clicked_1 = FILTER data BY was_clicked == 1;

-- TODO (C): Group the data by the appropriate field
display_filtered= GROUP clicked_1 BY display_site;

/* TODO (D): Create a new relation which includes only the
 *                display site and the total cost of all clicks
 *                on that site
 */
total_cost_grouped = FOREACH display_filtered GENERATE group AS display_site, SUM(clicked_1.cpc) AS total_cost;
--DUMP total_cost_grouped;

-- TODO (E): Sort that new relation by cost (ascending)
total_cost_grouped = ORDER total_cost_grouped BY total_cost ASC;

-- TODO (F): Display just the first three records to the screen
top_4 = LIMIT total_cost_grouped 4;
DUMP top_4;
```

Part 2
If group by one, got (PRESENT,165606)
(TABLET,106509)
(DUALCORE,95124), script is

```
data = LOAD 'dualcore/ad_data[1-2]/part*'
                AS (
                        campaign_id:chararray,
                        date:chararray,
                        time:chararray,
                        keyword:chararray,
                        display_site:chararray,
                        placement:chararray,
                        was_clicked:int,
                        cpc:int
                );

clicked_1 = FILTER data BY was_clicked == 1;
```

```
display_filtered = GROUP clicked_1 BY keyword;


total_cost_grouped = FOREACH display_filtered GENERATE group AS keyword, SUM(clicked_1.cpc)
AS total_cost;



total_cost_grouped = ORDER total_cost_grouped BY total_cost DESC;


top_3 = LIMIT total_cost_grouped 3;
DUMP top_3;
```

Part 3

```
-- Load only the ad_data1 and ad_data2 directories
data = LOAD 'dualcore/ad_data[1-2]/part*' AS (campaign_id:chararray,
                    date:chararray, time:chararray,
                    keyword:chararray, display_site:chararray,
                    placement:chararray, was_clicked:int, cpc:int);

-- Include only records where the ad was clicked
clicked = FILTER data BY was_clicked == 1;

-- A: Group everything so we can call the aggregate function
grouped = GROUP clicked ALL;

-- B: Count the records
total = FOREACH grouped GENERATE COUNT(clicked.was_clicked);


-- C: Display the result to the screen
DUMP total;
```

b) the result is 18243

``````````````````````````````````````````````````````````````````````````````````````````````````````````````````````````````
``````````````````````````````````````````````````````````````````````````````````````````````````````````````````````````````

Part 4

a)

```
data = LOAD 'dualcore/ad_data[1-2]/part*'

      AS (campaign_id:chararray,
                date:chararray, time:chararray,
                keyword:chararray, display_site:chararray,
                placement:chararray, was_clicked:int, cpc:int);

clicked = FILTER data BY was_clicked IS NOT NULL;


grouped = GROUP clicked ALL;


total = FOREACH grouped GENERATE MAX(clicked.cpc)*50000;

DUMP total;
```

b)

the estimated cost is (8000000)